Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, eds

SELECTING RANDOM LATIN HYPERCUBE DIMENSIONS AND DESIGNS THROUGH ESTIMATION OF MAXIMUM ABSOLUTE PAIRWISE CORRELATION

Alejandro S. Hernandez
Thomas W. Lucas
Paul J. Sanchez

Naval Postgraduate School
1 University Circle
Monterey, California 93940, USA

ABSTRACT

Latin hypercubes are the most widely used class of design for high-dimensional computer experiments. However, the high correlations that can occur in developing these designs can complicate subsequent analyses. Efforts to reduce or eliminate correlations can be complex and computationally expensive. Consequently, researchers often use uncorrected Latin hypercube designs in their experiments and accept any resulting multicollinearity issues. In this paper, we establish guidelines for selecting the number of runs and/or the number of variables for random Latin hypercube designs that are likely to yield an acceptable degree of correlation. Applying our policies and tools, analysts can generate satisfactory random Latin hypercube designs without the need for complex algorithms.

1 INTRODUCTION

Experimentation is fundamental to science and knowledge acquisition. In many cases, physical experimentation is infeasible due to safety, money, time, or resource constraints. Indeed, in a number of important areas—e.g., long-term effects of various policies on global climate, possible future military conflicts, or emergency response to large-scale nuclear accidents—comprehensive physical experiments are impractical. In situations lacking real-world experimental data, computer models are often instrumental in understanding these complex issues and in communicating possible consequences of policy options to decision makers.

Computer simulations used in the above areas may contain thousands of input variables and/or take a long time (even many days) to run (Kleijnen et al. 2005). Researchers have many techniques to extract information from these models. Among them are designs of experiments (DOEs) that are specifically developed for efficiently exploring high-dimensional computer models. The design specifies the inputs for the experiments. Given that n experiments are to be conducted over k continuous input variables, also known as factors, the DOE is specified as an $n \times k$ design matrix $\boldsymbol{X}$, where n and k are the design dimensions. Each column of $\boldsymbol{X}$ represents a factor and each row specifies a single design point as a particular combination of values for the set of factors. Of course, the quality of information obtainable by analyzing the data from the experiments depends critically on the design. For example, we cannot identify a nonlinear response for a quantitative input variable that has only two levels in the design.

If we know in advance what meta-models we desire to fit and the error structure of the experiments, then an optimal design may exist (Fedorov 1972). However, in many cases, especially in exploratory analysis, we desire designs that "allow one to fit a variety of models" (Santner, Williams, and Notz 2003, p. 124). For such situations, Latin hypercube (LH) sampling (McKay, Beckman, and Conover 1979) has proven to be an invaluable technique. In fact, LHs are reported to be the predominant design for

978-1-4673-4781-5/12/$31.00 ©2012 IEEE

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Selecting Random Latin Hypercube Dimensions and Designs through Estimation of Maximum Absolute Pairwise Correlation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,1 University Circle,Monterey,CA,93940**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the 2012 Winter Simulation Conference, 9-12 Dec, Berlin, Germany

14. ABSTRACT
Latin hypercubes are the most widely used class of design for high-dimensional computer experiments. However, the high correlations that can occur in developing these designs can complicate subsequent analyses. Efforts to reduce or eliminate correlations can be complex and computationally expensive. Consequently, researchers often use uncorrected Latin hypercube designs in their experiments and accept any resulting multicollinearity issues. In this paper, we establish guidelines for selecting the number of runs and/or the number of variables for random Latin hypercube designs that are likely to yield an acceptable degree of correlation. Applying our policies and tools, analysts can generate satisfactory random Latin hypercube designs without the need for complex algorithms.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**12**	

experiments involving computer simulations (Buyske and Trout 2001). With increasing frequency, simulation software packages—even spreadsheet simulation add-ons—can generate LHs (Sugiyama and Chow 1997). Furthermore, under general conditions, LH designs perform well in comparison to other common experimental design options (Johnson et al. 2008).

A key reason for the popularity of LHs is that they come with minimal restrictions on the number of factors and sampling budget. Moreover, LHs have good space-filling properties, i.e., they are good at providing "information about all portions of the experimental region" (Santner, Williams, and Notz 2003, p. 124). In addition, the resultant output data allow us to fit many different models to multiple outputs from a single experimental set. This flexibility extends to visual investigations of the data (Sanchez et al. 2012), as we get many viewpoints from which to observe the relationships between inputs and outputs.

Many analytical techniques that experimenters apply to computer outputs—such as regression modeling and partition trees—suffer when there is multicollinearity among the input variables (Montgomery, Peck, and Vining 2001; Kim and Loh 2003). Consequently, analysts usually desire a design matrix with a diagonal variance/covariance structure; i.e., zeros in the off-diagonal elements. Unfortunately, generating random LH designs can produce substantial correlation among the columns of the design matrix, especially when k is small and n is not much larger than k.

Many methods have been developed that reduce correlations among the columns of an LH. These often work quite well, especially when n is large relative to k (Hernandez 2008). However, they typically utilize sophisticated techniques or require specialty software. Thus, *uncorrected* random LHs remain in widespread use and analysts work with the inefficiencies that result from multicollinearity. Our research offers a framework to sensibly choose dimensions for an LH design and, prior to generating the design matrix, inform the scientist of the expected degree of multicollinearity in the experimental data.

The organization of the remainder of this paper follows. Section 2 describes random LH (RLH) generation and the possible occurrence of high correlations. It also introduces the maximum absolute pairwise correlation (ρ_{map}) as a key measure for discriminating between LH designs. Section 3 describes the behavior of ρ_{map} in relation to n and k, and presents parsimonious multiple linear regression models that predict the expected value of ρ_{map}, which can be realized from a collection of 200 RLH designs, given a specific design dimension. Section 4 extends Section 3's results by considering other numbers of RLH designs. We summarize our results in Section 5.

2 BACKGROUND

RLH generation is so named to emphasize the randomness in the construction of its columns. Our work is based on the ability to describe the degree of nonorthogonality we should expect from this randomness.

2.1 RLH Generation

Generating an RLH is relatively simple. In LH sampling, the input variables are treated as random variables with known distribution functions. For each input variable X^j, $j = 1, \ldots, k$, "all portions of its distribution [are] represented by input values" by dividing its range into "n strata of equal marginal probability $1/n$, and [sampling] once from each stratum" (McKay, Beckman, and Conover 1979, p. 240). In practice, and we will do so here, many analysts take a fixed value in each stratum (e.g., the median). In such a case, the design points all fall on a lattice (Patterson 1954). For each X^j, the n sampled input values are assigned at random to the n design points, with all $n!$ possible permutations being equally likely. This generates the X^j column in the design matrix. The permutation process is performed independently for each of the k input variables. Therefore, for each column X^j, all of the n input values appear exactly once in the design. Also, for a given row in the design matrix, all of the n^k potential combinations of the input variable values have an equal chance of occurring. A value in the jth column

and *i*th row is labeled X_i^j. Creating a lattice RLH corresponds to independently generating k permutations of the first n natural numbers and appropriately scaling the columns to cover the variables' ranges. A total of $(n!)^k$ designs exist (Joseph and Hung 2008).

In a sampling method in which all possible RLH designs are equally probable, the probability that a highly correlated design occurs can be large—especially for small n, and k close to n. For example, we generated 1000 4×3 RLH design matrices and measured each correlation. Over 77% of the designs have a correlation greater than 0.8 or less than -0.8, and nearly 25% have at least one pair of columns with perfect correlation. The likelihood of constructing highly correlated RLHs calls for a systematic way to select a suitable design dimension and obtain an uncorrected LH with acceptable nonorthogonality.

2.2 Measure of Nonorthogonality

We want to specify a measure that we can use to distinguish between unacceptable and acceptable RLHs. Owen (1994) and Tang (1998) recognize that assessing a design based on correlation is a reasonable way to obtain one with an acceptable degree of nonorthogonality. The correlation between any two column vectors, X^i and X^j, in a design is

$$\rho_{ij} = \frac{\sum_{b=1}^{n}[(X_b^i - \bar{x}^i)(X_b^j - \bar{x}^j)]}{\sqrt{\sum_{b=1}^{n}(X_b^i - \bar{x}^i)^2 \sum_{b=1}^{n}(X_b^j - \bar{x}^j)^2}}, \tag{1}$$

where $\bar{x}^i$ is the mean value of the elements of column i of the design matrix.

Among the $\binom{k}{2}$ pairwise correlations in a design with k variables, the pairwise columns with the largest magnitude can have the greatest impact on the meta-model derived from the experiment. We focus on the maximum absolute value of the pairwise correlation (ρ_{map}) to identify acceptable RLHs:

$$\rho_{map} = \max\left\{|\rho_{ij}|,\ \forall(i \neq j)\right\}. \tag{2}$$

Controlling the worst case, pairwise correlation bounds the degree of multicollinearity in the design.

2.3 Methods to Reduce or Eliminate Nonorthogonality

To reduce the correlation in LH designs, scientists use methods that apply a series of transformation procedures to change the original design. McKay, Beckman, and Conover (1979) started a revolution in experimental design by introducing Latin hypercube sampling (LHS) as a means to decrease the variance in the estimates derived from computer experiments. Studies to improve on the LHS design have taken scientists on different paths: transformation or column generation.

Iman and Conover (1982) developed a transformation matrix from the rank matrix associated with the design matrix as a means to control correlation. Florian (1992) used Cholesky's decomposition of the rank correlation matrix to derive a transformation matrix that reduces the correlation among the columns of the design's corresponding rank matrix. Owen (1994) used Gram-Schmidt orthogonalization (Leon, 2002) to produce a transformation matrix for the lattice version of the LH.

Other methods completely eliminate correlation during construction of the columns. Ye (1998) proposed orthogonal LHs (OLHs) as a new class of designs, developing OLH designs when the number of runs for an experiment is 2^m or $2^m + 1$, and the number of factors is $2m - 2$, for $m > 1$. Cioppa and Lucas (2007) modified construction of these designs to generate nearly orthogonal columns (with $\rho_{map} \leq$

0.03), thus increasing the number of factors that the design addresses without increasing the number of runs, and designated them nearly OLHs (NOLHs).

Steinberg and Lin (2006) rotated two-level factorial designs to construct OLHs for $n = 2^h$, with h a power of 2, and the maximum number of factors being $B_h \times h$, where $B_h = \lfloor (n-1)/h \rfloor$. For instance, for $n = 16$ runs, $h = 4$ and $B_h = 3$, so an O_{12}^{16} design is possible. Pang, Liu, and Lin (2009) showed that an OLH may be constructed for $n = p^d$, where p is a prime number and d is a power of 2. This generalized construction method includes Steinberg and Lin's approach (2006) as a special case ($p = 2$).

Hernandez (2008) used an optimization routine to generate an NOLH for almost any nonsaturated, run-variable combination, along with some saturated designs. The basis of this algorithm is a mixed integer program formulation.

A commonality among these methods is that they require specialized algorithms and are computer-intensive. Furthermore, some methods work for only relatively few values of n and k.

3 A NEW APPROACH

In this paper, we develop a methodology to create experimental designs that can address a variety of experimental challenges without any additional burden on resources. In lieu of complex algorithms, we seek a simplified alternative that leverages the ease of generating RLHs. If an RLH has acceptable correlation among its columns, an experimenter can reap the benefits that an efficient design offers, with a significant reduction in the computational cost or investment of time in developing the design. In practice, experimenters often generate many RLHs and select the best one for their experimentation. Our study develops tools based on Equation 2 to help analysts choose an appropriate design dimension. Secondly, analysts can set a threshold ρ_{map} to select acceptable designs.

3.1 Creating the $\overline{\rho_{map}^{\min}}$ Table

We begin our work with an initial set of data that consists of 42 (n, k) design combinations. We chose the 42 (n, k) pairs to correspond to known OLH and NOLH designs. Using Cioppa's (2002) dimensional convention, we explore combinations of $n = 2^m + 1$ for up to $m = 8$, and $k = m + \binom{m-1}{2}$ for up to $m = 16$. We initially examine design dimensions as small as $n = 17$, $k = 7$, and as large as $n = 257$, $k = 121$. We consider only those designs with $n > k$, i.e., those in which we can fit a main effects model.

The data to create our correlation tables is generated from 200 RLHs for each specific (n, k) combination and the associated ρ_{map} values. We use G to designate the number of RLHs from which to select our experimental plan (i.e., G200). From among the 200 RLHs, we take the one that has the minimum value for ρ_{map} and label it $\rho_{map}^{\min}$. We repeat this process 1,000 times for each (n, k) combination and examine the resulting $\rho_{map}^{\min}$ values. We find that the distribution of $\rho_{map}^{\min}$ appears to be roughly bell-shaped (i.e., reasonably well approximated by a normal distribution). Therefore, the table entry for each (n, k) combination is the average $\rho_{map}^{\min}$ from 1,000 trials: $\overline{\rho_{map}^{\min}}$. Since the collected data are a random sample from the population of RLHs for the specific (n, k) combination, we can use the resulting analysis to make general statements about that population.

Values of $\overline{\rho_{map}^{\min}}$ for different design dimensions vary, but the standard deviation for any given design dimension is relatively small, with the largest being 0.025 (See Figure 1). We see that the largest empirical deviation occurs for a small design ($n = 17$, $k = 16$), and the smallest standard deviation is for a

large RLH (n = 257, k = 106). Smaller LH designs usually present challenges in the degree of nonorthogonality among the matrix columns (Hernandez 2008).

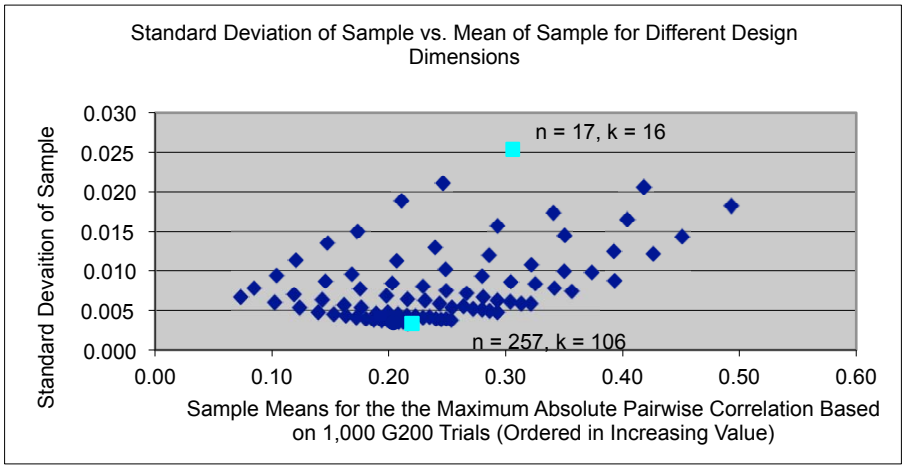

Figure 1: The standard deviation of $\rho_{map}^{\min}$ for 1,000, G200, is relatively small compared to its $\overline{\rho_{map}^{\min}}$ value. The largest standard deviation occurs in small designs, and the smallest deviations in the larger designs.

Table 1 is the complete set of $\overline{\rho_{map}^{\min}}$ G200 values from Hernandez (2008), and it includes 115 (n, k) combinations. It allows the experimenter to ascertain a realistic expectation of $\rho_{map}^{\min}$ for a given design dimension and within G trials. It also maps alternate design combinations for an RLH that may possess the ρ_{map} that the experimenter needs. If the table indicates that the initial design dimension is not likely to attain the desired ρ_{map} within 200 RLHs, then the table guides the experimenter to increase n, decrease k, increase G, or some combination of the above.

Table 1: The G200 table shows the $\overline{\rho_{map}^{\min}}$ for different design combinations.

		k															
	G= 200	7	11	16	22	29	37	46	56	67	79	92	106	121	137	154	172
	17	0.308	0.416	0.494													
	25	0.246	0.343	0.403	0.451												
	33	0.210	0.294	0.350	0.394	0.427											
	49	0.173	0.239	0.286	0.322	0.350	0.374	0.394									
n	65	0.148	0.207	0.248	0.280	0.304	0.326	0.342	0.356								
	97	0.121	0.169	0.203	0.230	0.248	0.266	0.280	0.293	0.304	0.313	0.322					
	129	0.104	0.147	0.177	0.198	0.216	0.231	0.244	0.254	0.264	0.272	0.280	0.286	0.293			
	193	0.085	0.119	0.144	0.161	0.177	0.189	0.199	0.208	0.216	0.223	0.230	0.235	0.240	0.245	0.249	0.253
	257	0.074	0.104	0.124	0.140	0.153	0.163	0.173	0.180	0.187	0.194	0.199	0.204	0.209	0.212	0.217	0.220
	513	0.052	0.072	0.088	0.099	0.109	0.116	0.122	0.128	0.133	0.137	0.141	0.145	0.148	0.151	0.153	0.156
	1025	0.037	0.051	0.062	0.070	0.077	0.082	0.087	0.091	0.094	0.097	0.100	0.102	0.104	0.107	0.109	0.111

Table usage is straightforward. Consider an analyst who wishes to explore 20 factors with a design that has a $\rho_{map} \le 0.20$. Table 1 shows that an acceptable design is likely to be found within 200 randomly generated LHs. It also frames the dimensions to the ranges $97 < n < 129$ and $16 < k < 22$. The analyst can then adjust the experimental design by increasing or decreasing the number of runs, factors,

and/or generated RLHs. However, this tabular guidance does not fully address the analyst's need. We develop another tool to more precisely specify the design dimension.

3.2 Developing a Function to Estimate Expected $\rho_{map}^{\min}$

We would like to use n and k to predict the expected value of $\rho_{map}^{\min}$ from 200 RLH designs. Our goal is a formula that is sufficiently simple to use in a calculator. Using a predictive formula allows the experimenter to find different (n, k) combinations that meet an acceptable correlation threshold.

We examine $\overline{\rho_{map}^{\min}}$ data from the original 42 n and k combinations (Hernandez 2008) to create the predictive function. Patterns are evident in the relationships between $\overline{\rho_{map}^{\min}}$ and (n, k) when either n or k is constant and the other changes. Grouping $\overline{\rho_{map}^{\min}}$ values, based on the number of sample runs, uncovers specific patterns in the data. The left-hand side of Figure 2 shows that while n is constant, the relationship between $\overline{\rho_{map}^{\min}}$ and k appears logarithmic. Similarly, grouping the number of factors (k) on the right-hand side chart indicates an exponentially decaying pattern between $\overline{\rho_{map}^{\min}}$ and n for constant k.

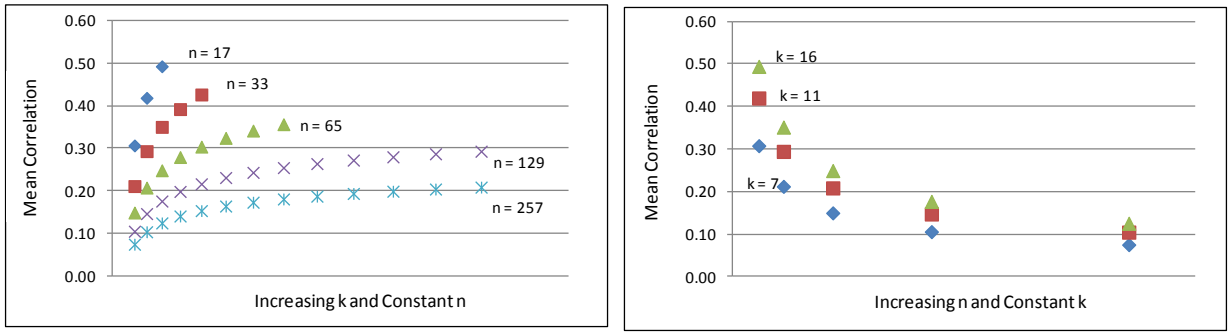

Figure 2: The left-hand chart shows a logarithmic pattern between the $\overline{\rho_{map}^{\min}}$ and k when n is constant. The right-hand chart indicates an exponentially decaying pattern appears between $\overline{\rho_{map}^{\min}}$ and n for constant k.

Transforming n and k shows nearly linear relationships with $\overline{\rho_{map}^{\min}}$. Owen (1994) established that the variance of the root mean square correlation (ρ_{rms}) of an uncorrected LH design is related to n^{-3}. However, Owen does not explicitly include k in his formulas. We examine different transformations of n to determine its linear relationship with $\overline{\rho_{map}^{\min}}$ and find that $n^{-2/3}$ has a near linear relationship with $\overline{\rho_{map}^{\min}}$ when k is constant, as shown on the left-hand side of Figure 3 for $k = 7$. Likewise, a transformation of k to $k^{-1/3}$ reveals a nearly linear relationship with $\overline{\rho_{map}^{\min}}$ when n is constant. The right-hand side of Figure 3 illustrates this relationship for $n = 257$.

Owen (1994) provides support for the exponentially decaying relationship between $\overline{\rho_{map}^{\min}}$ and n. He fit models for $k = n - 1$ to predict root mean square correlation (ρ_{rms}) for an RLH as a function of n, while we vary both n and k to predict $\overline{\rho_{map}^{\min}}$. Owen found the relationship to be:

$$\log(\rho_{rms}) \approx -0.5\log(n). \tag{3}$$

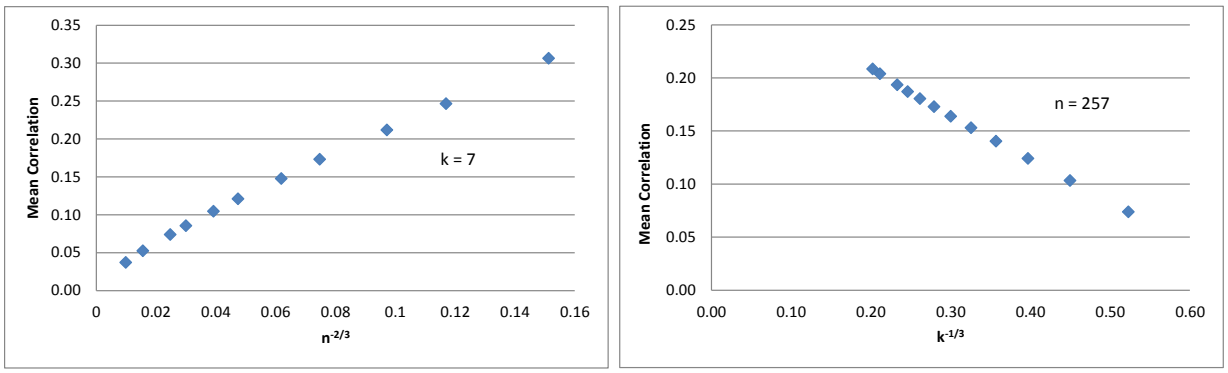

Figure 3: The left-hand chart shows a nearly linear relationship between $\overline{\rho_{map}^{\min}}$ and $n^{-2/3}$ when $k = 7$. The right-hand side has a similar relationship between $\overline{\rho_{map}^{\min}}$ and $k^{-1/3}$ when $n = 257$.

The preliminary exploration of the linear relationship between $\overline{\rho_{map}^{\min}}$ and $n^{-2/3}$, as well as $k^{-1/3}$, supports development of a multiple linear regression (MLR) model. Our exploration begins with a master simple linear regression (MSLR) model. The general MSLR model for $\overline{\rho_{map}^{\min}}$ regressed on $k^{-1/3}$ is:

$$\overline{\rho_{map}^{\min}} = \beta_0 + \beta_1 k^{-1/3} + \varepsilon \quad . \tag{4}$$

We group the data in terms of n and regress $\overline{\rho_{map}^{\min}}$ on $k^{-1/3}$ in each group to create an SLR model, designating each instance of n as SLR_n. Although the data sets are small, the coefficient of determination for each SLR_n model is greater than 0.99. From the set of SLR_n models, the estimated intercept and coefficient in Table 2 shows the change in coefficient values as n changes.

Table 2: Values of transformed n and corresponding SLR_n intercepts and coefficients.

n	$n^{-2/3}$	β_0	β_1
17	0.1513	1.0839	–1.4848
33	0.0972	0.7825	–1.0914
65	0.0619	0.5625	–0.7919
129	0.0392	0.4087	–0.5863
257	0.0247	0.2907	–0.4184

The left-hand side of Figure 4 illustrates a nearly linear relationship between $n^{-2/3}$ and the intercepts, while the right-hand side shows a linear relationship with the variable coefficients of SLR_n, thereby supporting the idea of developing a linear model.

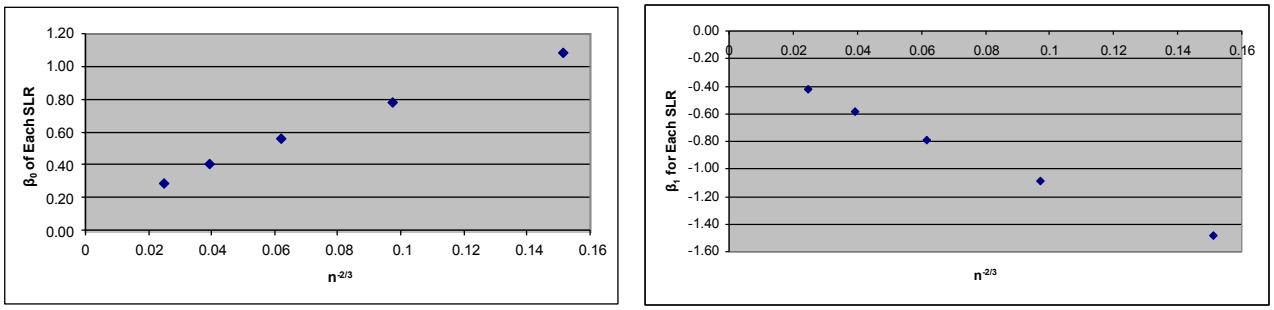

Figure 4: Nearly linear relationships between transformed n and β_0 of SLR_n models (left-hand chart) and transformed n and β_1 of SLR_n models (right-hand chart).

We develop SLR models in terms of $n^{-2/3}$ for the intercept, as well as the coefficient, in the general MSLR. We regress the intercepts from the set of SLR_n models onto corresponding $n^{-2/3}$ values and designate the resulting SLR model as SLR_{β_0}. Similarly, we regress variable coefficients from the set of SLR_n models onto corresponding $n^{-2/3}$ values and designate the model as SLR_{β_1}. These simple linear regression models define the MSLR in terms of k:

$$\text{MSLR} = SLR_{\beta_0} + SLR_{\beta_1} * k^{-1/3}. \tag{5}$$

Substituting the actual expressions for SLR_{β_0} and SLR_{β_1} into the MSLR and collecting terms for simplification, the resulting expression to estimate $\overline{\rho_{map}^{\min}}$ follows:

$$\left(\overline{\rho_{map}^{\min}}\right)^E = 0.161 + 6.206n^{-2/3} - 0.251k^{-1/3} - 8.328n^{-2/3}k^{-1/3}. \tag{6}$$

Notably, this preliminary study, based on 42 (n, k) combinations, identifies the need for an interaction term in the equation. Figure 5 shows a nearly linear relationship between the interaction term and $\overline{\rho_{map}^{\min}}$.

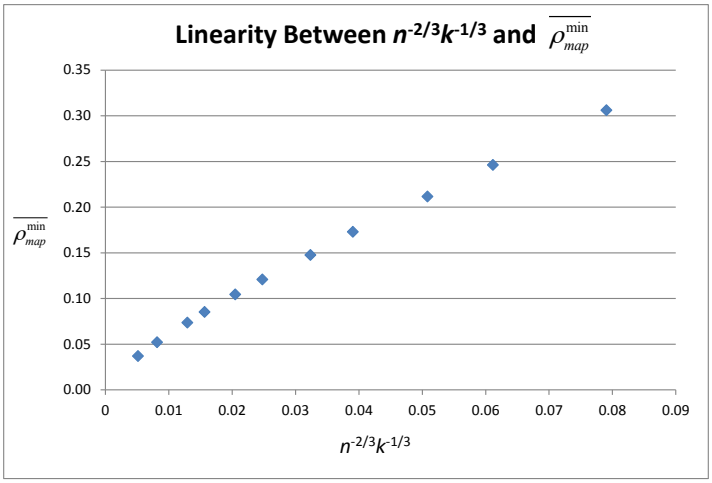

Figure 5: The interaction term of transformed n and k is nearly linear with $\overline{\rho_{map}^{\min}}$, indicating that the MLR model should have an interaction term. For clarity, we select $k = 7$ for this illustration.

Equation 6 is a tractable, compact model in its representation of n and k. We see that as n increases and k remains constant, the first term is dominant and the mean maximum absolute pairwise correlation decreases. As one may expect, for larger values of k we require much larger values of n to reduce correlation to the same level as smaller k.

We examined the adequacy of Equation 6 to predict $\overline{\rho_{map}^{\min}}$ using a larger set of 115 design combinations, including the data from the 42 initial design combinations. We concluded that an MLR, with two main terms and one interaction term, is sufficient to accurately predict $\overline{\rho_{map}^{\min}}$. Using least squares on all the data, we developed a new MLR that applies to the (n, k) ranges found in Table 1:

$$\left(\overline{\rho_{map}^{\min}}\right)^{E} = 0.0873 + 7.859n^{-2/3} - 0.109k^{-1/3} - 11.702n^{-2/3}k^{-1/3}. \quad (7)$$

The coefficients derived from least squares are understandably different from the combined SLR models. However, the polarity for each term is in sync with Equation 6. Equation 7 has an adjusted R-square (R_{adj}^2) of 0.978, indicating its adequacy as an estimator for $\overline{\rho_{map}^{\min}}$ (Montgomery, 2005). Given any two entries from among (n, k), and $\overline{\rho_{map}^{\min}}$, an analyst can easily solve for the other.

The residuals associated with the fit to Equation 7 show some curvature—suggesting a higher order model might fit better. Thus, we extended the model to include quadratic terms and possible interactions for transformed n and transformed k. It results in the following eight-term equation with $R_{adj}^2 = 0.999$.

$$\begin{aligned}\left(\overline{\rho_{map}^{\min}}\right)^{E} &= 0.0305 + 0.0321*k^{-1/3} - 0.1008*k^{-2/3} + 13.0684*n^{-2/3} - 68.3808*n^{-4/3} \\ &\quad - 30.1278*k^{-1/3}n^{-2/3} + 254.892*k^{-1/3}n^{-4/3} + 17.9311*k^{-2/3}n^{-2/3} - 254.839*k^{-2/3}n^{-4/3}.\end{aligned} \quad (8)$$

3.3 A Log Transform Regression Model for $\overline{\rho_{map}^{\min}}$

Examination of the G200 data for 115 design combinations suggests that log transformation of n and k, as well as $\overline{\rho_{map}^{\min}}$, can also be useful for predicting the expected value for $\rho_{map}^{\min}$. So, we also develop a model that includes variables log(n), log(k), k and the interaction of log(n) and log(k). The results show a definite linear relationship between log($\overline{\rho_{map}^{\min}}$) and the individual variables, to include the interaction term. The resulting model has an R_{adj}^2 of 0.993. All explanatory variables, as well as the intercept, are significant. Residual analysis shows the adequacy of the model and we accept it as a viable alternative:

$$\log\left(\overline{\rho_{map}^{\min}}\right) = -2.395 - 0.021k - 0.503\log(n) + 1.162\log(k) + 0.007\log(n)\log(k). \quad (9)$$

We remind the reader that Equation 6 was developed in an exploratory phase using a smaller set of data, and therefore Equations 7, 8, or 9 are preferable. The user can choose whichever of these models best suits their needs. We find Equation 7 attractive for our purposes—it is parsimonious, clearly shows the impact of n and k on correlation, and requires no logarithmic reinterpretation of the explanatory or response variables, all of which make it easy to use.

4. EQUATIONS AND TABLES FOR DIFFERENT VALUES OF *G*

The experimenter may not wish to generate or even consider G = 200 RLHs before selecting a suitable design. The manner in which the experimenter generates RLH designs may also be a constraint. To

alleviate such circumstances, we provide $\overline{\rho_{map}^{\min}}$ tables and $\left(\overline{\rho_{map}^{\min}}\right)^{E}$ expressions for different values of *G*. We introduce the symbol, $\left(\rho_{map}^{\min}\right)_{G}$, as the best maximum absolute pairwise correlation value in *G* trials and $\left(\overline{\rho_{map}^{\min}}\right)_{G}$ for the average of any number of sets of *G* trials. The corresponding formula to estimate the best maximum absolute pairwise correlation value in *G* iterations is designated as $\left(\overline{\rho_{map}^{\min}}\right)_{G}^{E}$.

Investigating the impact of different values of *G* reveals notable observations. Previous work shows that for $G > 200$ the values of $\rho_{map}^{\min}$ vary only slightly from trial to trial. Conversely, as *G* decreases, the variance in $\rho_{map}^{\min}$ is more pronounced. To retain utility to experimenters, we set the lower bound for *G* at 10 and develop equations for $G = \{10, 25, 50, 75, 100, 125, 150, 175, \text{and } 200\}$.

With some slight modifications, we use the same methodology as in Section 3 to explore the relationship of transformed *n* and *k* values, as well as their interaction term. We develop new MLR models through least squares for each *G*. The corresponding $\left(\overline{\rho_{map}^{\min}}\right)_{G}^{E}$ models in increments of 25, with the exception of the last model at $G = 10$, are listed below.

$$\left(\overline{\rho_{map}^{\min}}\right)_{G175}^{E} = 0.0873 + 7.864n^{-2/3} - 0.109k^{-1/3} - 11.682n^{-2/3}k^{-1/3} \tag{10}$$

$$\left(\overline{\rho_{map}^{\min}}\right)_{G150}^{E} = 0.0874 + 7.870n^{-2/3} - 0.108k^{-1/3} - 11.650n^{-2/3}k^{-1/3} \tag{11}$$

$$\left(\overline{\rho_{map}^{\min}}\right)_{G125}^{E} = 0.0875 + 7.872n^{-2/3} - 0.107k^{-1/3} - 11.611n^{-2/3}k^{-1/3} \tag{12}$$

$$\left(\overline{\rho_{map}^{\min}}\right)_{G100}^{E} = 0.0875 + 7.883n^{-2/3} - 0.106k^{-1/3} - 11.578n^{-2/3}k^{-1/3} \tag{13}$$

$$\left(\overline{\rho_{map}^{\min}}\right)_{G75}^{E} = 0.0877 + 7.886n^{-2/3} - 0.105k^{-1/3} - 11.502n^{-2/3}k^{-1/3} \tag{14}$$

$$\left(\overline{\rho_{map}^{\min}}\right)_{G50}^{E} = 0.0877 + 7.912n^{-2/3} - 0.103k^{-1/3} - 11.423n^{-2/3}k^{-1/3} \tag{15}$$

$$\left(\overline{\rho_{map}^{\min}}\right)_{G25}^{E} = 0.0881 + 7.945n^{-2/3} - 0.0988k^{-1/3} - 11.270n^{-2/3}k^{-1/3} \tag{16}$$

$$\left(\overline{\rho_{map}^{\min}}\right)_{G10}^{E} = 0.0883 + 7.996n^{-2/3} - 0.0902k^{-1/3} - 11.014n^{-2/3}k^{-1/3} \tag{17}$$

Coefficients for these models are similar. However, the magnitude of most correlation values makes the subtleties in each *G*-specific expression important. These formulas provide the experimenter an option for *G*, along with choices of (*n*, *k*) and ρ_{map}.

5 CONCLUSIONS

Use of LH designs to conduct simulation experiments is prevalent in academia, the Department of Defense, and industry. Efficient LH designs provide researchers with a valuable tool for isolating the impact of dominant factors on outputs of interest. However, multicollinearity in these designs complicates interpretation and affects accuracy of meta-models that come from the corresponding experiments. The body of work to reduce or eliminate correlations in LH designs is extensive. Historically, construction of these designs is computer intensive and time consuming, but these resources are not always available to an experimenter.

We simplify the process of constructing a design that meets a worst-case correlation threshold. We define ρ_{map} as a measure of nonorthogonality. Using this measure as a basis, we develop tools and present an approach to obtain designs with acceptable nonorthogonality through RLH generation for the (n, k) combinations spanned in Table 1 (n up to 1025 and k up to 172) for G between 5 and 200. Our research efforts enable analysts to obtain effective designs for their needs without specialized software programs or complex algorithms.

REFERENCES

Buyske, S., and Trout, R. 2001. Advanced Design of Experiments. Statistics 591 Lecture Series, Rutgers University.

Cioppa, T.M. 2002. Efficient Nearly Orthogonal and Space-Filling Experimental Designs for High-Dimensional Complex Models. Doctoral Dissertation, Monterey, CA: Naval Postgraduate School.

Cioppa, T.M., and Lucas, T.W. 2007. Efficient Nearly Orthogonal and Space-Filling Latin Hypercubes. *Technometrics*, 45-55.

Fedorov, V.V. 1972. *Theory of Optimal Experiments*. New York: Academic Press.

Florian, A. 1992. An Efficient Sampling Scheme: Updated Latin Hypercube Sampling. *Probabilistic Engineering Mechanics*, *7*, 123-130.

Hernandez, A.S. 2008. Breaking Barriers to Design Dimensions in Nearly Orthogonal Latin Hypercubes. Doctoral Dissertation, Monterey, CA, Naval Postgraduate School.

Iman, R.L., and Conover, W.J. 1982. A Distribution-Free Approach to Inducing Rank Correlation among Input Variables. *Communications in Statistics – Simulation Computations*, *11*(3), 311-334.

Joseph, V.R., and Hung, Y. 2008. Orthogonal-Maximin Latin Hypercube Designs. *Statistica Sinica*, *18*, 171-186.

Johnson, R.T., Montgomery, D.C., Jones, B., and Fowler, J.W. 2008. Comparing Designs for Computer Experiments. Proceedings of the 2008 Winter Simulation Conference, pp. 463-470.

Kim, L., and Loh, H. 2003. Classification Trees and Bivariate Linear Discriminant Node Models. *Journal of Graphical and Statistics*, *12*, 512-530.

Kleijnen, J.P.C., Sanchez, S.M., Lucas, T.W., and Cioppa, T.M. 2005. A User's Guide to the Brave New World of Designing Simulation Experiments. *INFORMS Journal on Computing*, *17*(3), 263-289.

Leon, S.J. 2002. *Linear Algebra with Applications*. Upper Saddle River, NJ: Prentice Hall.

McKay, M.D., Beckman, R.J., and Conover, W.J. 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, *21*(2), 239-245.

Montgomery, D.C. 2005. *Design and Analysis of Experiments*. Hoboken, NJ: John Wiley and Sons, Inc.

Montgomery, D.C., Peck, E.A., and Vining, C.G. 2001. *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons, Inc.

Owen, A.B. 1994. Controlling Correlations in Latin Hypercube Samples. *Journal of the American Statistical Association*: *Theory and Methods*, *89*(428), 1517-1522.

Pang, F., Liu, M.Q., and Lin, D.K.J. 2009. A Construction Method for Orthogonal Latin Hypercube Designs with Prime Power Levels. *Statistica Sinica*, *19*, 1721-1728.

Patterson, H.D. 1954. The Errors of Lattice Sampling. *Journal of the Royal Statistical Society*, *Series B (Methodological)*, *16*(1), 140-149.

Sanchez, S.M., Lucas, T.W., Sanchez, P.J., Nannini, C.J., and Wan, H. 2012. Designs for Large-Scale Simulation Experiments, with Application to Defense and Homeland Security. *The Design and Analysis of Computer Experiments.* In K. Hinkelmann. (Ed.) *Volume 3*: *Special Designs and Applications*, *pp. 413-441*. Hoboken, NJ: Wiley.

Santner, T.J., Williams, B.J., and Notz, W. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer Verlag.

Steinberg, D.M., and Lin, D.K.J. 2006. A Construction Method for Orthogonal Latin Hypercube Designs. *Biometrika*, *93*(2), 279-288.

Sugiyama, S.O., and Chow, J.W. 1997. @Risk, Riskview and BestFit. *OR/MS Today*, *24*(2), 64-66.

Tang, B. 1998. Selecting Latin Hypercubes Using Correlation Criteria. *Statisca Sinica*, *8*, 965-977.

Ye, K.Q. 1998. Orthogonal Column Latin Hypercubes and Their Application in Computer Experiments. *Journal of the American Statistical Association – Theory and Methods*, *93*(444), 1430-1439.

AUTHOR BIOGRAPHIES

ALEJANDRO S. HERNANDEZ is an Associate Professor in the Systems Engineering Department at the Naval Postgraduate School. He holds a B.S. in Civil Engineering from the United States Military Academy, an M.S. and Ph.D. in Operations Research from the Naval Postgraduate School, and a Masters in Strategic Studies from the United States Army War College. His email address is ahernand@nps.edu.

THOMAS W. LUCAS is a Professor in the Operations Research Department at the Naval Postgraduate School, where he has been a leader in advancing the development and use of simulation experiments and efficient design since 1998. In this regard, he has co-founded and is the Codirector of the Simulation Experiments and Efficient Designs Center. Professor Lucas has a Bachelor of Science in Industrial Engineering and Operations Research from Cornell University, a Master of Science in Statistics from Michigan State University, and a Doctorate in Statistics from the University of California at Riverside. His email address is twlucas@nps.edu.

PAUL J. SANCHEZ is a faculty member in the Operations Research Department at the Naval Postgraduate School. He has an SB in Economics from MIT and an M.S. and Ph.D. in Operations Research from Cornell University. He has a long-standing interest in applying design of experiments to simulation analysis. His email address is pjsanche@nps.edu.